

## Research Article

# Towards Prediction of Pancreatic Cancer Using SVM Study Model

Yushan Qiu<sup>1\*</sup>, Hao Jiang<sup>2</sup>, Kazuaki Shimada<sup>3</sup>, Nobuyoshi Hiraoka<sup>4</sup>, Kensei Maeshiro<sup>5</sup>, Wai-Ki Ching<sup>1</sup>, Kiyoko F Aoki-Kinoshita<sup>6</sup> and Koh Furuta<sup>7</sup>

<sup>1</sup>Department of Mathematics, The University of Hong Kong, Hong Kong

<sup>2</sup>Department of Mathematics, Renmin University of China, China

<sup>3</sup>Division of Hepatobiliary and Pancreatic Surgery, National Cancer Center Hospital, Japan

<sup>4</sup>Division of Pathology, National Cancer Center Research Institute, Japan

<sup>5</sup>Department of Surgery, Fukuoka University, Japan

<sup>6</sup>Department of Bioinformatics, Soka University, Japan

<sup>7</sup>Division of Clinical Laboratories, National Cancer Center Hospital, Japan

**\*Corresponding author**

Yushan Qiu, Department of Mathematics, The University of Hong Kong, Pokfulam Road, Hong Kong; Tel: +852-9799-1929; Fax: +852-2559-2225; Email: yushanqiu2526374@163.com

**Submitted:** 11 February 2014

**Accepted:** 25 March 2014

**Published:** 02 May 2014

**Copyright**

© 2014 Qiu et al.

**OPEN ACCESS****Keywords**

- Pancreatic cancer
- Feature selection
- Machine learning
- Prediction

**Abstract**

Pancreatic cancer is known to be a difficult disease to diagnose early, and early research mainly focused on predicting the survival rate of pancreatic cancer patients. The correct prediction of the various disease states can greatly benefit the patient and also assist the design of effective and personalized therapeutics. The issue of how to integrate the available laboratory data with classification techniques is an important and challenging research issue. In this paper, we proposed a useful approach to construct a feature space which serves as a significant predictor for classification. Furthermore, we developed a novel method to identify the outliers which are important for improving the classification performance. Using our preoperative clinical laboratory data and histologically confirmed pancreatic cancer samples, computational experiments are performed with the use of Support Vector Machine (SVM) to predict the status of the patients. We further tested the method by employing the Multi-Layer Perceptron (MLP) kernel with a three-fold cross-validation to assess the predictive power of the selected features. Experimental results on the prediction of cancer state of patients indicate that our method performs very well on pancreatic samples obtained in the clinical environment.

**ABBREVIATIONS**

SVM: Support Vector Machine; MLP: Multi-Layer Perceptron

**INTRODUCTION**

In recent decades, pancreatic cancer has received a lot of attention since the survival rate is extremely low with surgery as the only treatment. There is an urgent need for early diagnosis and treatment of pancreatic cancer to further improve survival rate. Currently, the various popular statistical studies have struggled to predict patient survival by analyzing the relationships between clinicopathological data and newly developed or found biomarkers. However, it is quite rare to identify practical biomarkers successfully by these types of studies since patient survival is a complicated issue which is related to many factors, such as environmental background, genetic background, age and tumor size, etc. Thus, the current new research direction is

to predict various histological characteristics of the tumor itself [1] rather than patient survival. Studies in [2,3] indicated that histological tumor differentiation is a strong predictor of the venous or lymphatic permeation for cancer cells and invasion patterns of colon cancers and gastric cancers. This suggests that histological tumor differentiation [4–6] and lymph node metastasis [7–11] could be a good predictor when designing therapeutic strategies for the common type pancreatic cancer. Therefore, it is worthwhile to evaluate these potential biological properties and provide predictive information of cancer cell behavior pre-operatively. Thus our work aims to predict

- (1) Lymph node metastasis and
- (2) Tumor metastasis in pancreatic cancer.

If we can predict or determine the tumor characteristics before any surgical or even non-surgical treatment, then we

can design an effective and personalized therapeutics before treatment so as to enhance the patient survival rate significantly. In [12], Inductive Logic Programming (ILP) technique has been proposed to generate some rules based on the background knowledge and the examples, which are expressed in first-order logic programming. ILP is a promising technique, which creates some valuable rules that reveal the significant features contributing to the diagnosis of pancreatic cancer [12]. The identified features give us a clue that we can regard them as the most significant predictors and further apply them to predict the status of the patients.

Classification of biological data is an important issue in clinical research. Furthermore, a large amount of laboratory data is available in the clinical laboratory, which can therefore provide a good source for investigation and prediction of various disease states. Researchers usually focus only on physical samples for various reasons, but a large amount of leftover laboratory data is available for re-utilization [13,14]. We propose that one can take advantage of not only the leftover physical samples but also leftover laboratory data. In this study, we analyze the accumulated leftover laboratory data which involves many features. Various clinical test results of a particular patient reflect his/her past biological condition. Thus we hypothesize that the clinical tests in general can be considered as a  $n$ -dimensional pathological feature set, where  $n$  refers to the number of clinical tests. By taking the  $n$  samples and grouping them by a particular status, this data can be used to train a classifier, which can then be assessed by the cross-validation technique.

In this work, we use a support vector machine (SVM) to perform classification. SVMs have been popular in many research field for classification due to their ability to handle multi-dimensional data and to pinpoint specific features that are important for classification. Furthermore, the trained SVM can be used to make predictions given the clinical data of a new patient. From the perspective of improving the classification performance, removing irrelevant predictors could dramatically improve the classification performance and achieve a much more accurate prediction [15,16]. Furthermore, according to the “ugly duckling” theorem [17], it makes different classes more similar after adding irrelevant predictors and hence results in degrading classification performance. In view of this, feature selection could play a vital role in improving the performance of classification. Therefore we propose a novel approach to construct the feature space that involves the significant features. Furthermore, we propose a novel approach to identify the outliers in the data set and then exclude them when analyzing the data. Outlier detection can be treated as a part of the data preprocessing. A study in [18] has shown that outlier detection can efficiently and correctly detect the high dimensional nonlinear outlier features and has considerable practical value. Therefore, it is very important to determine the outliers and exclude them in the model. In this work, we propose an efficient approach to identify the outliers and we apply the SVMs model in predicting the disease states. Our results on the prediction of cancer state of patients indicate that our method performs very well on pancreatic samples obtained in the clinical environment compared to other well known models. Therefore we show that there is a great potential for such mathematical models to be used in clinical applications.

## MATERIALS AND METHODS

### Materials

In the study set, 174 surgically resected and histologically confirmed [19,20] common type pancreatic cancer cases at the National Cancer Center Hospital in Japan are utilized for our analysis. Based on the pathological reports information, tumor differentiation status and lymph node metastatic status (N0; negative nodal metastasis or N+; positive nodal metastasis), are used as the basis of classification. We prepare two data sets based on the above classification criteria: Set Diff (tumor differentiation between poorly differentiated vs. others) and the Set N1 (N+ vs. N0). For Set Diff, poorly to moderately differentiated tumor samples are considered as positive samples (41), well differentiated tumor samples are taken as negative samples (133). As for Set N1, we define N0 as negative findings while and the rest as positive, with a total of 86 positive samples. We use the following clinical data laboratory data from the same cancer cases: CEA, CA19-9, Glucose, Elastase I, Serum Amylase, C-reactive protein (CRP), Serum Glucose (GLU), Fibrin degradation product (FDP), Fibrinogen (FIBG) and Antithrombin III (ATIII). We also use data regarding age, sex, tumor location, tumor size (TS mm), number of lymphocytes (LymphNum) and lymphocyte ratio (LymphCell).

### Methodology

**Feature selection:** To identify the important features, we perform feature ranking using several available feature selection criteria, including “entropy”, “t-test”, “ROC”, and “bhattacharyya”. These criteria assess the significance of each feature for separating the two labeled groups. Entropy computes the Kullback-Leibler distance or divergence [21] and t-test computes the absolute value two-sample t-test with pooled variance estimate [22]. ROC calculates the empirical receiver operating characteristic (ROC) curve and the random classifier slope [23]. The Bhattacharyya criterion is based on the minimum attainable classification error [24].

After feature ranking, the following features are consistently extracted among the top four features for the Set N1: CA19-9, CEA, Elastase I, TS (mm). FIBG is selected twice among the five criteria. For Set Diff, the following features rank highly: selected all four times are CA19-9 and GLU, and three times are Elastase I and CEA. Other highly ranked features are Serum Amylase, FIBG, CRP, and FDP, as well as LymphCell. Table 1 provides an overview of the comparatively important features. It is interesting to note the both the data sets have several common highly ranked features: FIBG, CA19-9, CEA and Elastase I. This gives us a clue that we shall focus on these four features for classification.

**Feature discretization:** Inspired by [12], all the features used for analysis are initially converted into several groups: “low”, “normal” and “high”. The classification criterion is based on the definition of the normal ranges provided by National Cancer Center Hospital, described in detail in [12]. In order to quantitize the features, we adopt the discrete numbers: -1, 0 and 1 to indicate the low group, normal range, and high group, respectively. Therefore, we can construct feature vectors whose entries belong to the set  $\{-1, 0, 1\}$ .

**Table 1:** The Top Ranking Features for the 2 Data Sets.

Dataset	Rank	Entropy	T-test	ROC	Bhattacharyya
N1	1 <sup>st</sup>	CEA	TS	TS	CA19-9
	2 <sup>nd</sup>	FDP	CA19-9	CA19-9	Elastase I
	3 <sup>rd</sup>	CA19-9	CEA	CEA	FIBG
	4 <sup>th</sup>	GLU	FIBG	Elastase I	GLU
	5 <sup>th</sup>	LymphNum	GLU	FIBG	TS
Diff	1 <sup>st</sup>	CRP	CRP	CEA	Elastase I
	2 <sup>nd</sup>	CA19-9	GLU	Elastase I	CA19-9
	3 <sup>rd</sup>	GLU	CA19-9	CA19-9	FIBG
	4 <sup>th</sup>	LymphNum	Elastase I	LymphCell	Amylase
	5 <sup>th</sup>	CEA	CEA	GLU	GLU

**Abbreviations:** CRP: C - Reactive Protein; GLU: Serum Glucose; FDP: Fibrin Degradation Product; FIBG Fibrinogen; TS: Tumor Size; LymphNum: Number of Lymphocytes; LymphCell: Lymphocyte Ratio.

N1, Diff: Two different data sets based on the classification criteria. For Set Diff, poorly to moderately differentiated tumor samples are considered as positive samples, well differentiated tumor samples are taken as negative samples and for another Set N1, we define N0 as negative findings while set the others as positive ones.

**Grouping patients status:** In order to utilize the SVM to perform classification, we also prepared the associated label for each instance. Note that the stage of pancreatic cancer can be assessed by the tumor differentiation and lymph nodal metastasis which are used to describe the pancreatic cancer stages. Thus they somehow have strong correlation and the label should be consistent. In other words, we initially define the associated label for the data, i.e., for Set Diff, poorly to moderately differentiated tumor samples are considered as positive samples and well differentiated tumor samples are taken as negative samples. For the other Set N1, we define N0 as negative findings while the others are positive. To build a robust model, we can select the sub-instances where the labels overlap among these two data sets. As a result, 59 data instances are utilized to train the SVM. Integrating the four significant features we have selected, we could create a data set of size 59-by-4 (59 patients and 4 features).

**SVM model setting:** The mechanism of a SVM is to solve the maximum margin classifier for a given empirical data. The trial data is then classified based on the function constructed by the support vectors. SVM is an efficient hyperplane-learning algorithm. The optimal hyperplane can be solved by consider the following:

$$\max_{w \in H, b \in R} \min \{ \|x - x_i\| : x \in H, \langle w, x_i \rangle + b = 0, i = 1, \dots, m \}$$

Where  $H$  is the feature space and  $R$  refers to the space of the real values. To construct the optimal hyperplane, the following problem needs to be solved:

$$\min_{w \in H, b \in R} \tau(w) = \frac{1}{2} \|w\|^2$$

subject to  $y_i(\langle w, x_i \rangle + b) \geq 1, i = 1, \dots, m$ , where  $m$  is the number of samples. Therefore the decision function can be written in the form:

$$f(x) = \text{sgn}(\sum_{i=1}^m y_i \alpha_i (\langle x, x_i \rangle + b))$$

Where  $y_i$  is the label of the  $i$ th training sample,  $\alpha_i$  is the corresponding coefficient obtained from training process, and  $b$  is the bias through solving

$$\alpha_i [y_i (\langle w, x_i \rangle + b) - 1] = 0, i = 1, \dots, m.$$

Here we choose to use a multilayer perceptron (MLP) kernel which derives from neural network theory. Moreover, despite being only conditionally positive definite, it has been found to perform well in practice. The kernel function for the MLP takes the following form:

$$K(x, x_i) = \tanh(kx^T x_i + \theta)$$

Where  $k > 0$  and  $\theta < 0$ . The defaults values of  $k$  and  $\theta$  are 1 and -1, respectively. It is interesting to note that a SVM model using a sigmoid kernel function is equivalent to a two-layer, perceptron neural network. We remark that the MLP maps a feature vector not only from the original  $d$ -dimensional feature space, but from an intermediate implicit Hilbert feature space in which kernels calculate inner products. The kernel replaces the usual inner product between the weight vectors and the input vector (or the feature vector of the hidden layer). The objective is to boost the generalization capability of this universal function approximator. Moreover, the reason we adopt the MLP kernel is that the performance of the MLP kernel outperforms the other kernels including the radial basis function (RBF) and polynomial kernel in terms of the accuracy in this work. We are able to improve the classification performance for certain kernel types and their intrinsic parameters for the majority of the data sets.

**Performance evaluation:** Cross-validation is a popular technique for assessing the generalization ability of a statistical analysis. It holds the promise to evaluate the accuracy of a predictive model in practice. In particular, one round of cross-validation involves partitioning a sample of data into complementary subsets called training set and testing set respectively. The training set is the one used for the analysis while the testing set is used for validation. Usually multiple rounds of cross-validation are adopted with different partitions and the final validation result is obtained by taking the average over all the rounds.

**3-Fold cross validation:** The performance of our classification model is measured through 3-fold cross-validation, which entails using a single set of observations from the original sample as the validation data whilst the remaining observations are used as the training data. Generally speaking, in each round,

the original data is randomly divided into three subsamples. Among the three subsamples, a single subsample is retained as the validation data for testing the model. The remaining two subsamples are used as training data. The cross-validation process is then repeated three times (the folds), with each of the three subsamples used exactly once as the validation data. The average of the three results from the different folds is then used as a single estimation. The accuracy of the learning performance is assessed as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Here *TP* is the true positive number, *TN* is the true negative number, *FP* is the false positive number and *FN* is the false negative number. We note that accuracy is exactly the ratio of the correctly predicted items to the whole number of data instances. And for the sensitivity, it is measured as follows:

$$Sensitivity = \frac{TP}{TP + FN}$$

We remark that the sensitivity is the likelihood of the positive test, given that the candidate is a cancer patient. Also, we give the formula for the specificity:

$$Specificity = \frac{TN}{TN + FP}$$

It is noted that the specificity is exactly the probability of the negative test given that the candidate is not a cancer patient.

## Comparison to the state-of-the-art models

**Decision tree:** Decision tree learning is a method commonly used in data mining. It employs a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. In these tree structures, leaves represent classifications and branches represent conjunctions of features that lead to those classifications.

**K-Nearest neighborhood:** The K-nearest neighbor algorithm is the simplest method among all machine learning algorithms. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its K nearest neighbors (K is a positive integer and is typically small). If K = 1, then the object is simply assigned to the class of its nearest neighbor.

## RESULTS AND DISCUSSION

In this section, we shall discuss the classification performance of our model when compared with other machine learning techniques. In particular, the results have shown that our model is superior to the other typical techniques. Table 2 presents the accuracy comparison among the three approaches: Decision Tree Algorithm, K-Nearest Neighborhood method and our proposed method. Results are obtained by taking the average over the three rounds. Here we utilize the average of the accuracy to measure the performance which can be interpreted as the overall capability of the methods in prediction. We can see that for the average accuracy, our method performs better than the Decision Tree Algorithm and KNN method, reaching 70%. In terms of average specificity, our method is comparatively inferior, reaching at 70%. However, despite the slight inferiority of our method in average specificity, its value is actually good enough to be acceptable.

**Table 2:** Accuracy, Sensitivity and Specificity for the 3 methods.

Methods	Mean-AC	Mean-SE	Mean-SP
DT	65%	37%	79%
KNN	69%	1%	99%
SVM	70%	70%	70%

**Abbreviations:** DT: Decision Tree; KNN: K-Nearest Neighborhood; SVM: Support Vector Machine. Mean-AC, Mean-SE and Mean-SP represent the average of the accuracy, sensitivity and specificity for the pancreatic cancer profiles prediction, respectively.

Furthermore, taking into consideration of the average sensitivity, our proposed model can also achieve 70%, which evidently highlights the superiority of our method in fulfilling the task of prediction. Note that we just conduct the analysis on the average values; it is possible that the values are fragmented which implies some numbers are very large while some of them can be extremely slow. Thus there is a need to analyze the variance of the values. Standard deviation is a measure of how spreads out numbers are. We calculate the standard variance of the 100 (repeated the 3-fold cross-validation 100 times) values for SVM model. The corresponding results are 0.047, 0.091, 0.067 for accuracy, sensitivity and specificity, respectively which illustrates that our model is robust and stable. Apart from that, our proposed method makes it possible to achieve the balance between the sensitivity and the specificity since for other methods, they can reach comparatively high specificity while slightly sacrificing the sensitivity. Thus we have constructed a useful and effective framework to predict the patient status from the perspective of the classification performance.

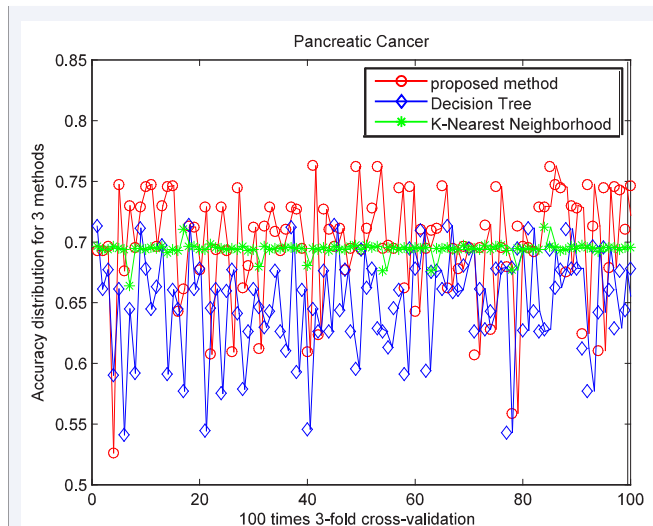
Furthermore, Figures 1-4 depict a more explicit explanation of the accuracy, sensitivity, specificity and AUC prediction performance for the 3 methods respectively. Figure 1 shows the accuracy distribution for the Decision Tree method, KNN method and SVM. We can see clearly that in general, the accuracy values for SVM are larger than the other two state-of-the-art methods. Furthermore, the minimum accuracy for SVM is only 53% which means the worst prediction accuracy is still better than average. This further confirms the superiority and effectiveness of our developed model in prediction. We were also concerned if our proposed method can outperform the other machine learning techniques using other criteria. Therefore, we conducted a study on the comparison of SVM with the other two state-of-the-art methods by employing sensitivity criterion. From the performance in Figure 2, one can easily see the superiority of SVM in classification. The sensitivity values are dramatically and consistently greater than the other two methods which confirm the reliability of SVM. Although the specificity of SVM is relatively lower than the other two methods, see for instance in Figure 3, the average of the specificity can also achieve 70% (described in Table 1), which is considered quite high for this field. The effectiveness of our method is also evaluated through comparing the AUC values. We can see from Figure 4 that the majority of AUC values are larger than the other two methods, which further confirms the effectiveness of our developed method for classification.

We also investigated the performance of SVM as a classifier for making predictions given a new data set. We employed the

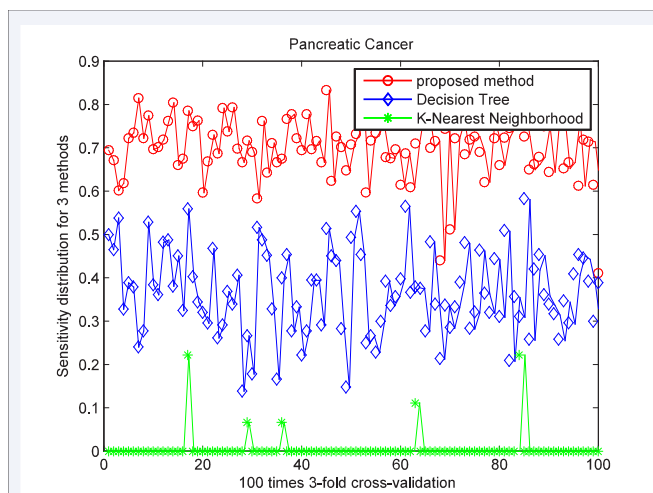


proposed model to predict new pancreatic cancer samples which were also provided by the National Cancer Center Hospital. The main idea was to utilize the already trained SVM to predict the status of a new patient. For our results in classification performance, the accuracy was also 72%, which is quite remarkable.

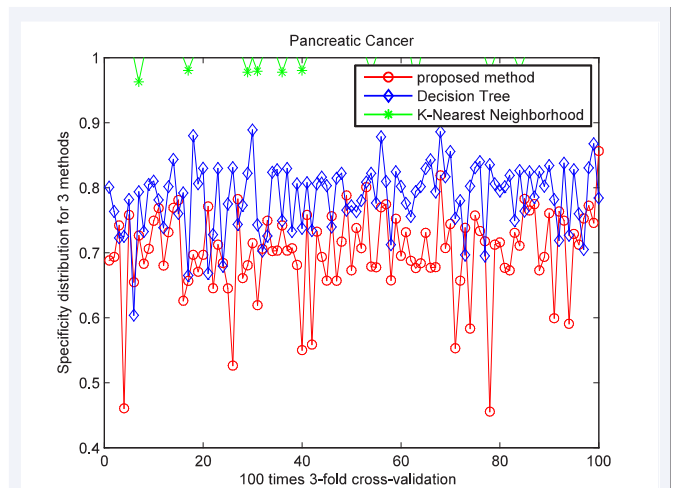
From the perspective of clinical diagnosis, the features we selected play a vital role in predicting the biological behavior. In this study, FIBG, CA19-9, CEA and Elastase I are used as the predictors in SVM. FIBG suggests that the coagulation system is somehow related to the lymph node metastasis. Furthermore, the characterization of both CA19-9 and CEA implies that their glycobiological structures may play a role in differentiation. The fact is that both CA19-9 and CEA have advantages in differential diagnosis between pancreatic cancer and chronic pancreatitis, assisting the assessment of treatment response, follow-up of pancreatic cancer and prognosis. CA19-9 is one of the more reliable tumor markers in pancreas cancers and authors in [25] proposed that CA19-9 is the serum tumor marker which



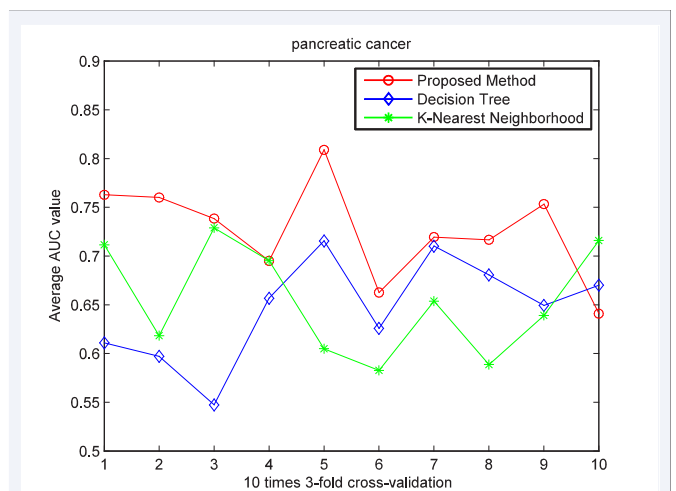
**Figure 1** Accuracy Distribution Comparisons for 3 methods.



**Figure 2** Sensitivity Distribution Comparisons for 3 methods.



**Figure 3** Specificity Distribution Comparisons for 3 methods.



**Figure 4** AUC value comparisons for 3 methods.

can assist improving the survival rate if the patient could be identified earlier. Additionally, Elastase I is a serine protease. It is synthesized by the acinar cells of the pancreas along with other digestion enzymes, thus it is an indicator of exocrine gland function of pancreas [26]. Furthermore, Elastase I can be a diagnostic clue for detecting pancreatic cancer [27]. Thus we claim that our analysis reflects the biological behavior of pancreatic cancer well.

Using the features from this work, and the SVM trained in this work, it should be possible to apply it to the clinical laboratory. Thus we plan on developing a software tool which will make it easier for clinicians to use this model to assist in the prediction of lymph node metastasis and tumor metastasis in pancreatic cancer.

## CONCLUSION

In this study, we present novel approaches for feature selection and outliers identification. They both contribute to improve the classification performance. Furthermore, we present a mathematical model that can directly utilize clinical laboratory

data for diagnosis based on multiple markers. The method we proposed here has a potential application for other diseases.

## ACKNOWLEDGEMENT

Research supported in part by GRF Grant, HKU CERG Grants, HKU Hung Hing Ying Physical Research Grant, National Natural Science Foundation of China Grant Nos. 11271144 and S201201009985. And it is supported by Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China.

## Author contribution

Jiang and Qiu came up with the idea. Jiang and Qiu designed the research plan. Qiu conducted the numerical experiments. Qiu and Furuta analyzed the results. Qiu, Ching and Aoki-Kinoshita wrote up the paper. Shimada provided clinical data. Maeshiro supervised the study. Hiraoka provided pathology data.

## REFERENCES

1. Wang Y, Xia XQ, Jia Z, Sawyers A, Yao H, Wang-Rodriquez J, et al. In silico estimates of tissue components in surgical samples based on expression profiling data. *Cancer Res.* 2010; 70: 6448-6455.
2. Chung CK, Zaino RJ, Stryker JA. Colorectal carcinoma: evaluation of histologic grade and factors influencing prognosis. *J Surg Oncol.* 1982; 21: 143-148.
3. Berti Riboli E, Secco GB, Lapertosa G, Di Somma C, Santi F, Percivale PL. Colorectal cancer: relationship of histologic grading to disease prognosis. *Tumori.* 1983; 69: 581-584.
4. Lim JE, Chien MW, Earle CC. Prognostic factors following curative resection for pancreatic adenocarcinoma: a population-based, linked database analysis of 396 patients. *Ann Surg.* 2003; 237: 74-85.
5. Kuhlmann KF, de Castro SM, Wesseling JG, ten Kate FJ, Offerhaus GJ, Busch OR, et al. Surgical treatment of pancreatic adenocarcinoma; actual survival and prognostic factors in 343 patients. *Eur J Cancer.* 2004; 40: 549-558.
6. Tobita K, Kijima H, Dowaki S, Oida Y, Kashiwagi H, Ishii M, et al. Thrombospondin-1 expression as a prognostic predictor of pancreatic ductal carcinoma. *Int J Oncol.* 2002; 21: 1189-1195.
7. Yeo CJ, Abrams RA, Grochow LB, Sohn TA, Ord SE, Hruban RH, et al. Pancreaticoduodenectomy for pancreatic adenocarcinoma: postoperative adjuvant chemoradiation improves survival. A prospective, single-institution experience. *Ann Surg.* 1997; 225:621-633.
8. Pawlik TM, Gleisner AL, Cameron JL, Winter JM, Assumpcao L, Lillemoe KD, et al. Prognostic relevance of lymph node ratio following pancreaticoduodenectomy for pancreatic cancer. *Surgery.* 2007; 141: 610-618.
9. Zacharias T, Jaeck D, Oussoultzoglou E, Neuville A, Bachelier P. Impact of lymph node involvement on long-term survival after R0 pancreaticoduodenectomy for ductal adenocarcinoma of the pancreas. *J Gastrointest Surg.* 2007; 11: 350-356.
10. Hattangadi JA, Hong TS, Yeap BY, Mamon HJ. Results and patterns of failure in patients treated with adjuvant combined chemoradiation therapy for resected pancreatic adenocarcinoma. *Cancer.* 2009; 115: 3640-3650.
11. Katz MH, Wang H, Fleming JB, Sun CC, Hwang RF, Wolff RA, et al. Long-term survival after multidisciplinary management of resected pancreatic adenocarcinoma. *Ann Surg Oncol.* 2009; 16: 836-847.
12. Qiu YS, Shimada K, Hiraoka N, Maeshiro K, Ching WK, Aoki-Kinoshita, et al. Knowledge discovery for pancreatic cancer using inductive logic programming. *Proceeding of the 7th International Conference on Systems Biology (ISB 2013)*, Huangshan, China, 2013.
13. Gotoh M, Nakatani T, Masuda T, Mizuguchi Y, Sakamoto M, Tsuchiya R, et al. Prediction of invasive activities in hepatocellular carcinomas with special reference to alpha-fetoprotein and des-gamma-carboxyprothrombin. *Jpn J Clin Oncol.* 2003; 33: 522-526.
14. Mitsunaga S, Kinoshita T, Hasebe T, Nakagohri T, Konishi M, Takahashi S, et al. Low serum level of cholinesterase at recurrence of pancreatic cancer is a poor prognostic factor and relates to systemic disorder and nerve plexus invasion. *Pancreas.* 2008; 36: 241-248.
15. Agrawal R, Gehrke J, Gunopulos D, Raghavan P. Automatic subspace clustering of high dimensional data for data mining applications. *Proceedings of the 1988 ACM-SIGMOD International Conference on the Management of Data (SIGMOD98)*, Seattle, WA, 1988: 94-105.
16. Dy J, Brodley CE. Feature subset selection and order identification for unsupervised learning. *Proceeding in the seventeenth International Conference on machine Learning*, Stanford, CA, USA. 2000: 247-254.
17. Watanabe S. *Knowing and Guessing: A Quantitative Study of Inference and Information*. New York: U.S.A; 2007.
18. Peng XQ, Chen J, Shen HY. Outlier detection method based on SVM and its application in copper-matte converting. *Proceeding in Control and Decision Conference.* 2010; 628-631.
19. Sobin L, Gospodarowicz M, Wittekind C. *TNM Classification of malignant tumors*. 7<sup>th</sup> edn. Hoboken, NJ: John Wiley & Sons, Inc. 2009.
20. Hruban RH, Boffeta P, Hiraoka N. *World health organization classification of tumors. Pathology & Genetics. Tumors of the Digestive System*. 4th edn. Lyon: IARC Press 2010.
21. Dash M, Liu H. Feature selection for clustering. *Proc. of PAKDD-00* 2000: 10-121.
22. Su Y, Murali TM, Pavlovic V, Schaffer M, Kasif S. RankGene: identification of diagnostic genes based on expression data. *Bioinformatics.* 2003; 19: 1578-1579.
23. Chen XW, Wasikowski M. Fast: a roc-based feature selection metric for small samples and imbalanced data classification problems. *Proceeding in the 14th ACM SIGKDD International conference on knowledge discovery and data mining*, 2008. 124-132.
24. Xuan GR, Chai PQ, Wu MH. Bhattacharyya distance feature selection. *Pattern Recognition* 1996. *Proceedings of the 13th International Conference.* 1996; 2:195-199.
25. Rückert F, Pilarsky C, Grützmann R. Serum tumor markers in pancreatic cancer-recent discoveries. *Cancers (Basel).* 2010; 2: 1107-1124.
26. Burtis CA, Ashwood ER, Bruns DE, Sawyer BG. *Tietz Fundamentals of Clinical Chemistry*. 6<sup>th</sup> edn. In: St.Louis: Saunders Elsevier; 2008.
27. Hamano H, Hayakawa T, Kondo T. Serum immunoreactive elastase in diagnosis of pancreatic diseases. A sensitive marker for pancreatic cancer. *Dig Dis Sci.* 1987; 32: 50-56.

## Cite this article

Qiu Y, Jiang H, Shimada K, Hiraoka N, Maeshiro K, et al. (2014) Towards Prediction of Pancreatic Cancer Using SVM Study Model. *JSM Clin Oncol Res* 2(4): 1031.